

General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.
- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.
- This document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

7.9-10133
CR-158142

LARGE AREA CROP INVENTORY EXPERIMENT (LACIE)

N79-18398

Unclas
00133

Under NASA sponsorship
early and wide dis-
semination and without liability
made thereof.



NASA NOAA USDA

ACCURACY ASSESSMENT REPORT PHASE 1A NOVEMBER/DECEMBER 1974

(E79-10133) LARGE AREA CROP INVENTORY
EXPERIMENT (LACIE). ACCURACY ASSESSMENT
REPORT PHASE 1A, NOVEMBER - DECEMBER 1974
(Lockheed Electronics Co.) 44 p HC A03/MF
A01 CSCL 02C G3/43



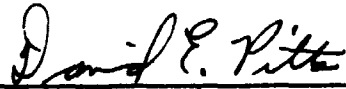
National Aeronautics and Space Administration
LYNDON B. JOHNSON SPACE CENTER

Houston, Texas
June 1976

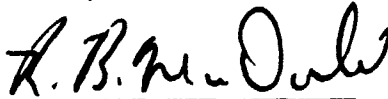
ACCURACY ASSESSMENT REPORT PHASE IA

NOVEMBER/DECEMBER 1974

Approved by:



D. E. Pitts



R. B. MacDonald

FOREWORD

Phase IA of LACIE was a period of operation that existed before the first classification system meeting the LACIE requirements was delivered and prior to the development of validated CAMS procedures. The purpose of this report is to provide an initial indication of the early LACIE system design performance in a few critical areas.

During the November/December 1974 operation, the Landsat I multispectral scanner data were retrospectively ordered for an area in central and southwestern Kansas. With respect to the biostages utilized in LACIE Phase I, segments per biostage were ordered as follows: biostage IA, 3 segments; biostage IB, 6 segments; biostage II, 5 segments; biostage III, 3 segments; and biostage IV, 11 segments.

In five intensive test sites, multitemporal data were received, which allowed the accuracy of classification to be evaluated for each biostage. The data received from the LACIE Phase IA LACIE Operation Report were: for sites 1106 and 1034, biostages I, II, III, and IV¹ were received; for

¹During LACIE Phase IA, biostages were defined as:

- I - Sowing to jointing
- II - jointing to soft dough
- III - soft dough to harvest
- IV - harvest to 36 days after harvest

site 1114, biostages II and IV were received; for site 1042, biostages I, III, and IV were received; and for site 1111, biostages III and IV were received.

The accuracy assessment activity involved comparisons of automatic data processing classification, analyst interpreter findings, and acreage estimation results to other available information, such as ground-truth data from SRS enumerative sites and the intensive test sites. These comparisons were also made to the official Statistical Reporting Service estimates for Kansas for the 1973-1974 crop year, and statistical analysis was performed on the data thus obtained.

ACKNOWLEDGEMENTS

This document represents the work of personnel from Lockheed Electronics Company, Inc., Aerospace Systems Division, Houston, Texas, and National Aeronautics and Space Administration under contract NAS9-12200.

CONTENTS

Section	Page
1.0 <u>SUMMARY</u>	1-1
2.0 <u>LACIE ACCURACY ASSESSMENT - NOVEMBER/</u> <u>DECEMBER OPERATIONS</u>	2-1
2.1 CAS AGGREGATION	2-1
3.0 <u>ERROR ANALYSIS - NOVEMBER/DECEMBER</u> <u>LACIE OPERATIONS</u>	3-1
3.1 LACIE DATA ANALYSIS	3-1
3.1.1 <u>Method</u>	3-2
3.1.2 <u>Field Data</u>	3-2
3.1.2.1 <u>Wheat error variance</u>	3-3
3.1.2.2 <u>Non-wheat error variance</u>	3-6
3.1.3 <u>ITS Wheat Percentage Data</u>	3-7
4.0 <u>ESTIMATION OF VARIANCE COMPONENTS OF LACIE</u> <u>ACREAGE ESTIMATES - NOVEMBER/DECEMBER 1974</u> <u>OPERATION</u>	4-1
4.1 LACIE DATA ANALYSIS	4-1
4.2 ESTIMATION OF THE SAMPLING ERROR COMPONENT	4-2
4.3 VARIANCE COMPONENTS	4-6
5.0 <u>REFERENCES</u>	5-1
APPENDIX A	A-1
APPENDIX B	B-1

FIGURE

Figure		Page
3-1	Comparison of actual percent wheat and LACIE computed percent wheat for five ITS in Kansas (correlation coefficient = 0.52).	3-8

TABLES

Table		Page
2-I	SAMPLE SEGMENTS FOR NOVEMBER SYSTEM	2-5
2-II	CRD WHEAT ACREAGE ESTIMATES AND THEIR MEASURES OF ACCURACY	2-7
4-I	ANALYSIS OF VARIANCE - WHEAT PROPORTION PER 1-SQUARE-MILE AREA (SRS SEGMENT) FOR FIVE INTENSIVE TEST SITES IN KANSAS	4-5
A-I	ANALYSIS OF VARIANCE - WHEAT FIELD DATA	A-4
A-II	ANALYSIS OF VARIANCE - NONWHEAT FIELD DATA	A-5
A-III	MAIN COEFFICIENTS FOR THE MODEL	A-6
A-IV	ITS WHEAT PERCENTAGE DATA	A-7
A-V	ANALYSIS OF VARIANCE - ITS WHEAT PERCENTAGES	A-8
A-VI	COEFFICIENTS IN THE MODEL	A-9

SECTION 1.0

SUMMARY

The results of the accuracy assessment activity for Phase IA of LACIE indicated that

1. the 90/90 criteria could be reached if the degree of accuracy of the LACIE performance¹ in Kansas could be equaled in other areas.
2. the classification of both wheat and nonwheat fields is significantly accurate for the three ITS segments analyzed in detail. The wheat field classification accuracy varied significantly for the segments. However, this was not so with respect to nonwheat fields.
3. biophase as well as its interaction with segment location turned out to be a significant factor for the classification performance. Analyst interpretation of segments for training the classifier was a significant error-contributing factor in the estimation of wheat acreage at both the field and the segment levels.

¹Only 28 of the 84 Kansas segments were processed, and these were limited to six crop reporting districts, only two of which had most of the segments that were allocated. Therefore, this was a limited test over a major wheat-producing region, and the accuracy of extrapolating these results to the entire central plains, including more marginal areas, would be questionable.

4. low percentages of wheat were overestimated and that high percentages of wheat were underestimated. (Regression showed a correlation coefficient, r , of .52 between actual and estimated percentages of wheat over five Kansas intensive test sites.
5. classification variances were approximately 3.6 times larger than the sample variance.

SECTION 2.0

LACIE ACCURACY ASSESSMENT - NOVEMBER/DECEMBER OPERATIONS

2.1 CAS AGGREGATION

In a CRD that is composed of only Group I and Group III type counties, the acreage estimate is computed using the following formula:

$$\hat{Y} = \left[1 + \frac{W_O}{W} \right] \sum_{k=1}^m A_k \hat{P}_k \quad (2-1)$$

where

$\hat{P}_k$ = Computed LACIE estimate of wheat proportion in k^{th} county.

A_k = Total acreage in k^{th} county reported in 1969.

W = Total wheat acreage during 1969 in m counties for which LACIE estimates are computed.

W_O = Total wheat acreage during 1969 in the remaining counties, say $(n - m)$, for which no LACIE estimates are computed.

The variance of $\hat{Y}$ is given by

$$S^2 = \left[1 + \frac{W_O}{W} \right]^2 \sum_{k=1}^m A_k^2 S_k^2 \quad (2-2)$$

where S_k^2 denotes the variance of estimate $\hat{P}_k$. (Observe that if $\hat{P}_k$ is a biased estimate, then S_k^2 should be considered the mean square error (MSE) for $\hat{P}_k$, and thereby one can obtain the MSE instead of the variance of $\hat{Y}$ by computing S^2 .) In present computations we shall assume that $\hat{P}_k$ is unbiased for all counties.

Next, for the collection of all six CRD's, a wheat acreage estimate and its variance is obtained by

$$\hat{X} = \sum_{i=1}^6 \hat{Y}_i \quad (2-3)$$

and

$$\sigma^2 = \sum_{i=1}^6 S_i^2 \quad (2-4)$$

where

$\hat{Y}_i$ = LACIE estimate for i^{th} CRD.

S_i^2 = Variance of $\hat{Y}_i$.

To compute the variance in equation (2-2), and thereafter in equation (2-4), the variances within counties, S_k^2 , $k = 1, 2, \dots, m$, need to be computed. Since a single segment is used in obtaining $\hat{P}_k$ for the k th county,

$k = 1, 2, \dots, m$, no direct estimate of S_k^2 can be made. However, an estimate of S_k^2 , $k = 1, 2, \dots, m$, can be achieved in the following indirect way.

From the 28 segments processed (table 2-I), the average wheat proportion computed is 0.326, and an estimate of the variance between segments is 0.01488. On the other hand, using the 1969 census data for the 28 counties for which the segments were processed, the variance between county wheat proportions is 0.00489. Now, considering that the variance between segments is composed of the variance within counties and the variance between counties, the variance within counties, a measure of sampling error, is obtained by the difference ($0.01488 - 0.00489 = 0.00999$).

However, if necessary, as is the case presently, it needs to be corrected with respect to the bias associated in estimating the average wheat proportion of all 28 counties. An estimate of this bias is $0.326 - 0.30 = 0.026$, where 0.30 is the SRS estimate for 1973-74. Accordingly, an estimate of S_k^2 is $0.00999 - 0.026^2 = 0.0093$. Now, replacing S_k^2 by 0.0093 in equation (2-2), an estimate of S^2 is obtained. Similarly, one obtains an estimate of σ^2 in equation (2-4).

Presented in table 2-II are each CRD wheat acreage estimate and an estimate of its standard deviation, coefficient of variation (CV), bias, and the confidence level for the bias

to be within plus or minus 10 percent of the actual wheat acreage. Bias is calculated in terms of the difference between a LACIE estimate and its corresponding SRS estimate for the 1973-1974 crop year. The formula for computing the confidence level α is given by

$$\alpha = \Phi\left(\frac{0.1 - k}{cv}\right) - \Phi\left[-\left(\frac{0.1 + k}{cv}\right)\right] \quad (2-5)$$

where $cv = \frac{s}{Y}$

s = standard deviation

$$B = E(\hat{Y}) - Y$$

$E(\hat{Y})$ = expected acreage estimate

Y = true acreage

$$K = \frac{B}{Y}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

TABLE 2-I.- SAMPLE SEGMENTS FOR NOVEMBER SYSTEM

State or region	CRD	County or site name	Site coordinates	Site number	Aggregated
Kansas	1	Graham	N39:29 - W099:55	1018	X
Kansas	2	Osborne	N39:32 - W099:00	1101	X
Kansas	2	Ottawa	N39:11 - W097:35	1102	X
Kansas	2	Rooks	N39:11 - W099:07	1103	X
Kansas	4	Lane	N38:25 - W100:26	1026	
Kansas	4	Ness	N38:36 - W099:47	1028	X
Kansas	4	Scott	N38:28 - W100:46	1029	X
Kansas	4	Trego	N39:02 - W099:49	1030	
Kansas	5	Barton	N38:37 - W098:50	1104	X
Kansas	5	Dickinson	N33:51 - W099:11	1105	
Kansas	5	Ellis ^a	N33:50 - W099:13	1106	X
Kansas	5	Ellsworth	N38:37 - W096:13	1107	X
Kansas	5	Lincoln	N39:09 - W098:14	1108	X
Kansas	5	Marion	N38:32 - W097:08	1109	X
Kansas	5	McPherson	N33:22 - W097:23	1110	X
Kansas	5	Rice ^a	N33:17 - W098:13	1111	X
Kansas	5	Rush	N38:33 - W093:05	1112	
Kansas	5	Russell	N39:05 - W093:50	1113	
Kansas	5	Saline ^a	N33:52 - W097:28	1114	X
Kansas	7	Hodgeman	N33:12 - W098:41	1039	
Kansas	7	Kearny	N37:56 - W101:17	1040	X

^aIntensive Test Sites

TABLE 2-I.- Concluded

State or region	CRD	County or site name	Site coordinates	Site number	Aggregated
Kansas	7	Morton ^a	N37:16 - W101:54	1042	X
Kansas	7	Seward	N37:20 - W100:38	1043	X
Kansas	7	Stanton	N37:43 - W102:01	1044	X
Kansas	7	Stevens	N37:15 - W101:06	1045	X
Kansas	8	Edwards	N37:45 - W099:23	1115	X
Kansas	8	Pawnee	N38:07 - W099:03	1116	X
Kansas	8	Reno	N38:09 - W097:49	1117	X
Kansas	8	Reno	N37:54 - W098:24	1118	
Kansas	8	Stafford	N38:02 - W098:40	1119	X
Kansas	7	Finney ^a	N38:04 - W101:02	1034	X
Kansas	7	Grant	N37:43 - W101:20	1036	X
Kansas	7	Gray	N37:48 - W100:38	1037	X
Kansas	7	Hamilton	N38:03 - W101:38	1038	X
Kansas	7	Haskell	N37:29 - W101:00	1065	X
Colorado	9	Baca	N37:29 - W102:48	1012	

^aIntensive Test Sites

TABLE 2-II.- CRD WHEAT ACREAGE ESTIMATES AND THEIR MEASURES OF ACCURACY

CRD	No. of counties	No. of segments processed	Wheat area estimate, $\hat{Y}$ (10 ³ acres)	Standard deviation, S (10 ³)	Coefficient of variation (C.V.), $100(S/\bar{Y})$	Relative difference, $\frac{(\hat{Y} - Y)}{\bar{Y}}$	Confidence level, $b_{100\alpha}$	Y^a
1	8	1	939	565	60	-0.23	15.1	1 217.1
2	11	3	1 358	287	21	.056	33.7	1 286.4
4	9	2	1 092	286	26	-.16	27.6	1 305.4
5	11	8	2 238	194	8.7	.27	6.2	1 759.8
7	14	10	2 438	250	10	.20	20.1	2 027
8	13	4	2 024	334	16.5	-.27	7.8	2 765.6
Sum	66	28	10 089	834	8.3	-0.026	76.3	10 361.8

^a Y = 1973-74 SRS estimate, Kansas Farm Facts 1974-1975, Kansas Department of Agriculture.

^b α = Probability that the difference $(\hat{Y} - Y)$ is within plus or minus 10 percent of Y .

SECTION 3.0

ERROR ANALYSIS - NOVEMBER/DECEMBER LACIE OPERATIONS

3.1 LACIE DATA ANALYSIS

Five intensive test site (ITS) segments located in Kansas were acquired and processed during the November-December 1974 Large Area Crop Inventory Experiment (LACIE) operation. For three of these segments, number 1034 in Finney, number 1042 in Morton, and number 1106 in Ellis Counties, multispectral scanner (MSS) data processing was completed for four biological phases. For the remaining two segments, number 1111 in Rice and number 1114 in Saline Counties, data processing was done for two biological phases. Two photointerpretations of each segment for its available biophases were completed by two randomly selected Analyst Interpreters (AI's). As a result, two sets of training data were used to obtain two different classification runs for each biophase of a segment.

For the classification and acreage performance evaluation, approximately 20 test fields, 10 wheat and 10 nonwheat fields, were randomly selected from the ITS area inside a segment, and the correctly classified pixels for each of these fields were determined under the Classification and Mensuration Subsystem (CAMS) operation. Also computed were percentages of wheat within an ITS area for each classification run.

Thus, the wheat acreage performance was evaluated at two levels: field size and ITS area size.

3.1.1 Method

The analysis of covariance technique is employed to analyze the classification and acreage performance data. This technique determines the contribution by each factor in the variance observed in the response measurements, in the presence of different covariable values. Considering a linear relationship between the response variable (e.g., LACIE-computed wheat acreage) and the covariable (e.g., actual wheat acreage), it regresses the former on the latter. It accounts for the variation in the response variable caused by the variation in the covariable. It computes the residual variation, total variation less that due to the covariable, and further provides a breakdown by various factors and any interactions between them, including that due to natural variation, called the error component.

3.1.2 Field Data

Three intensive test site segments (1034, Finney; 1106, Ellis; 1042, Morton) which were observed for all four biophases are considered for data analysis. Separate data analyses are made

for wheat and nonwheat fields performance data. Pairing the LACIE-computed acreage (response variable), wheat or nonwheat as the case may be, with the actual acreage (covariable) for each selected field, the variance observed in the LACIE-computed acreages was analyzed in terms of variations caused by the field size, segment location, biophase, and photo-interpretation. Details on the underlying design model, numerical values of main coefficients in the model, and the analysis of variance tables are given in Appendix A.

3.1.2.1 Wheat error variance.— In the case of wheat, the total variance depicted in the data is 2848.66, and the residual variance after accounting for the field size variation is 1307.93. So the variation in the field size accounts for 54 percent of the total variance:

$$\frac{2846.66 - 1307.93}{2848.66} \times 100 = 54$$

The error variance evaluated after accounting for the factors of segment location, biophase, photointerpretation, and factor interactions is 730.79, implying that all these factors and their interactions account for 44 percent of the residual variance:

$$\frac{1307.93 - 730.79}{1307.93} \times 100 = 44$$

or, 20 percent of the total variance:

$$\frac{1307.93 - 730.79}{2848.66} \times 100 = 20$$

This leaves 56 percent of the residual variance or 26 percent of the total variance unaccounted for.

In the absence of knowledge about any other factor influencing the performance, the leftover portion provides an estimate of natural/inherent variation in the system. Thus, in the present case, the system error at the field level is estimated to be 27 (i.e., $\sqrt{730.788}$) pixels, or roughly 30 acres in terms of standard deviation. By no means can this be regarded low, and, hence, the system needs improvement for it to be considered effective and successful.

As for the significance of different factors, the classification and acreage performance is significantly different from one biophase to another, but no such evidence exists in the case of segments. However, none of the biophases is uniformly better or worse when all segments are considered together.

This implies that the system performance depends simultaneously upon the segment location and the biophase. This could be expected because the effect of these two factors is dependent upon the types of competing crops and their state of growth in a segment location. And so, if either the competing crops

or their growth differs from segment to segment, the system performance can be expected to behave as shown presently. Next, there is a strong evidence of significant variation in the system performance due to different photointerpretations of a segment. This means a significant difference exists in training the classifier for automatic data processing classification and can be due to several reasons.

- A. One AI photointerprets the fields better than the other.
- B. Equal performance of AI's in terms of correct identification, but different sets of training fields are selected.
- C. Different types of bias in selecting training fields by AI's; e.g., one AI might be inclined to select a larger portion of a field than the other, or they might pick up a different number of training fields for various crops in a segment.

However, this is an indication that the inherent overlap in the MSS data of different crops is not too large to remain unaffected by AI performance; and given a better set of training fields, classification performance could be improved considerably. Refer to the technical report, Chhikara (March 1975), "Effect of Analyst Misinterpretation of Training Data

on Classification Performance," LEC-4353, which throws light on such theoretical aspects of this problem.

3.1.2.2 Non-wheat error variance.- In the case of nonwheat, the total variance obtained is 2914.60, and the residual variance is 581.12. Since

$$\frac{2914.60 - 581.12}{2914.60} \times 100 = 80$$

the variation in field size accounts for 80 percent of the total variance. Next, the error variance is 463.95, implying all other factors and their interactions account for 20 percent of the residual variance:

$$\frac{581.12 - 463.95}{581.12} \times 100 = 20$$

or, 4 percent of the total variance:

$$\frac{581.12 - 463.95}{2914.60} \times 100 = 4$$

This means that 80 percent of the residual variance or 16 percent of the total variance remains unaccounted for. Once again, this percentage needs to be reduced for the system to have an acceptable performance.

In this case, segment location, biophase, and interaction between the two are all significant factors, which means that the system performance is very much dependent upon what

segment is processed and for what biophase. However, the factor of AI is not significant. This could be due to merging all competing crops into one category - nonwheat.

3.1.3 ITS Wheat Percentage Data

The actual wheat percentage and the LACIE-computed wheat percentage for each ITS area in a segment processed for different biophases are given in table A-IV of appendix A. Altogether there are 30 response measurements, and these are not arranged in a design that can support an extensive statistical analysis. So any data analysis for this set of measurements will be limited in scope. Often, a much larger data set is needed for a proper and valid statistical inference, particularly when the system generating the data could cause large variation. As such, considering only the biophase factor and the actual wheat percentage (i.e., the covariable, which also represents the segment location factor in the model), an analysis of variance is presented in table A-V. The coefficient values for the terms in the model are presented in table A-VI of Appendix A. The total variance in data is 276.74, and 21 percent of it is accountable by the variation in actual wheat proportion in an ITS. The residual variance is 217.78, and most of it remains unaccounted for, implying that the factor for biophase is not significant compared to the inherent system variability (see fig. 3-1). In terms of standard deviation,

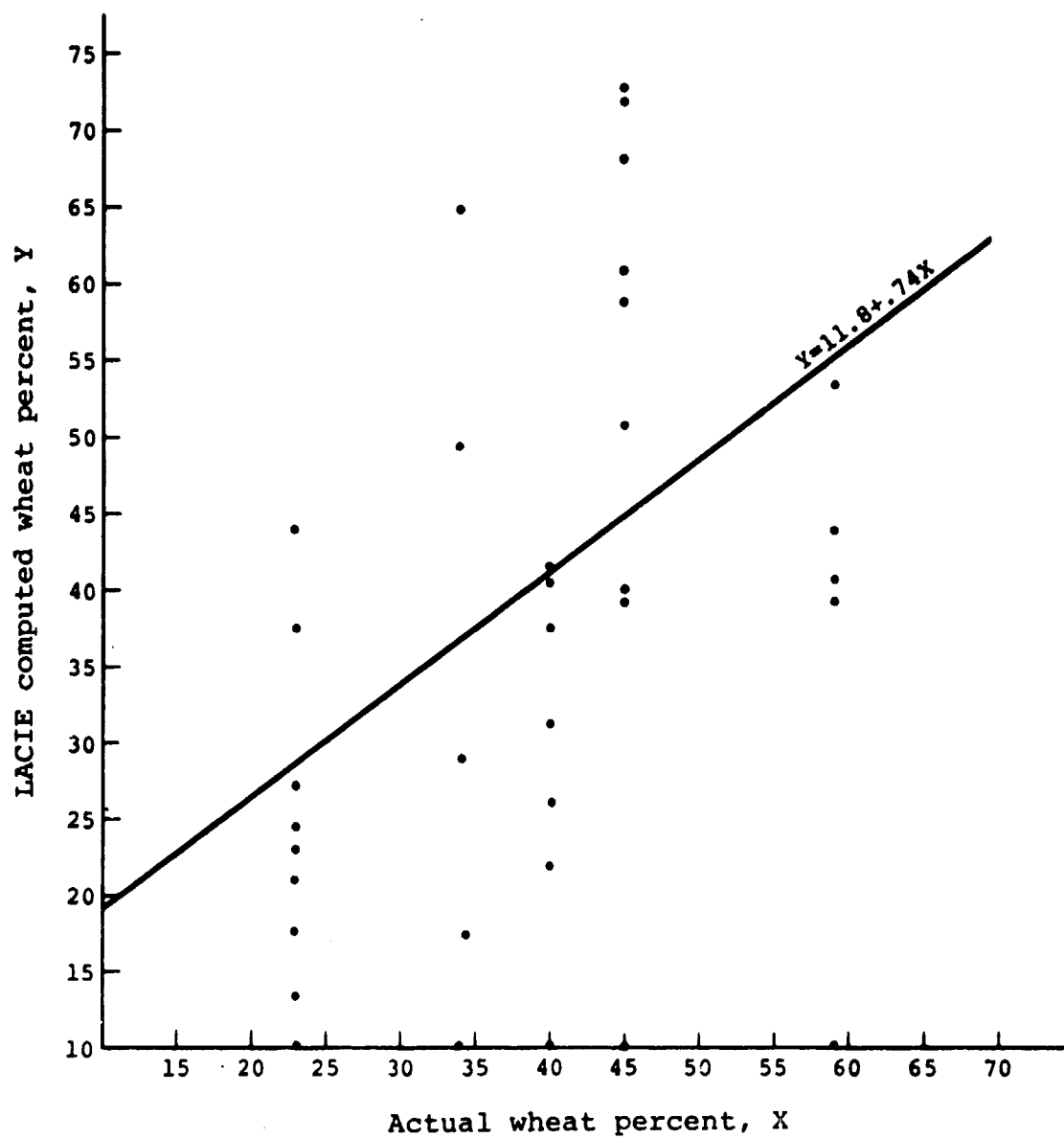


Figure 3-1.- Comparison of actual percent wheat and LACIE computed percent wheat for five ITS in Kansas (correlation coefficient = 0.52).

the latter is estimated to be 14.8 percent. Besides having this large a variation, there is an indication of an overall bias in estimating the wheat acreage. As pointed out earlier, this sort of conclusion is very tentative, and a larger set of experimental data is needed for a definitive study.

SECTION 4.0

ESTIMATION OF VARIANCE COMPONENTS OF LACIE ACREAGE

ESTIMATES - NOVEMBER/DECEMBER 1974 OPERATION

4.1 LACIE DATA ANALYSIS

The LACIE acreage estimate of a crop reporting district or state is subject to both the classification errors and the sampling errors. So, the variance of a LACIE estimate consists of two components: (1) variance due to classification errors, and (2) variance due to sampling error. In order to analyze the LACIE system performance, it is desirable to separate these two variance components and to evaluate their impact individually on the reliability of an acreage estimate.

During the LACIE-IA phase (November/December 1974 operation) segments acquired and processed were from Group I counties only. Acreage estimates were made for 28 counties, each containing a segment for which the wheat proportion was calculated, and 6 CRD's. In assessing these estimates for accuracy and reliability, their variances were estimated in section 1. Since these were all Group I counties, the variance estimates were obtained using within county variance, which was assumed to be the same for all the counties in the six CRD's. As such, the problem of variance decomposition reduces to obtaining a break down of the LACIE system variance into the two components mentioned in the previous paragraph.

4.2 ESTIMATION OF THE SAMPLING ERROR COMPONENT

In a communication (ref. 7), Fred Warren (affiliated with the Statistical Reporting Service, USDA) reported an estimate of 0.011983 for the within county variance for the SRS segments (average size = 823.4 acres) in Kansas. This was obtained using actual wheat proportions of 460 SRS segments distributed over 97 counties in 9 CRD's. Since the LACIE segment is larger in area (25426.5 acres), a smaller within county variance is expected because the actual wheat proportion for the LACIE segment will be less variable than that for the SRS segment in the county. Thus, the figure of 0.011983 is an upper bound for the within county variance for the LACIE segment. On the average, there are 30 SRS segments having the same area as 1 LACIE segment. If there is no correlation among wheat proportions of 30 SRS segments making up each LACIE segment, then selection of 1 LACIE segment in a county is equivalent to a random selection of 30 SRS segments in the county. In this case, ignoring classification errors, an estimate of within county variance for the LACIE segment is $0.011983/30 = 0.000394$. Since a certain amount of intra-segment correlation (i.e., the correlation among SRS segment

wheat proportions in a LACIE segment) is expected, this estimate of 0.000394 is, in fact, a lower bound for the within county variance for the LACIE segment. Thus, an unbiased estimate of the within county variance for the LACIE segment, which is the sampling error component since no classification errors are considered here, is expected to fall in between 0.000399 and 0.011983.

If the intra-segment correlation coefficient r is determined with respect to SRS segments, then an estimate of the sampling error variance component for the LACIE segment can be obtained as (e.g., see ref. 5, p. 242)

$$S_L^2 = \left(1 - \frac{n}{N}\right) \frac{[1 + (M - 1) r]}{nM} S_w^2 \quad (4-1)$$

where, in the case of Group I counties:

$n = 1$, the number of LACIE segments selected in a county.

$N = 21.5$, the average number of LACIE segments in a county.

$M = 30$, the average number of SRS segments having the same area as one LACIE segment.

$S_w^2 = 0.011983$, within county variance for the SRS segment as obtained in reference 7.

To determine a value of r , one needs to know the actual wheat proportions for the SRS segments clustered in different LACIE segments in a county. In general, this information is not available. However, an estimate of r is obtained in the following way.

Each of the five intensive test sites (ITS's) in Kansas was partitioned into 1-square-mile areas, and the wheat proportion for each of these areas was determined. For these wheat proportion data, an analysis of variance is presented in table 4-I. Based on this analysis, an estimate of r is obtained equal to 0.15. (See Appendix B for details.)

Now, making substitutions in equation (4-1), one gets

$$S_L^2 = \left[1 - \frac{1}{21.5} \right] \frac{1 + 29(0.15)}{30} (0.011983) = 0.00204 \quad (4-2)$$

Hence, the sampling error component (i.e., within county variance when no classification errors are made) for the LACIE segment is estimated to be 0.00204.

This estimate is derived on the basis of wheat proportion data from five ITS's, of which two were 5- by 6-square-mile areas and three were 3- by 3-square-mile areas. From the reliability viewpoint, this is a small number of samples. For example, the estimate of 0.00204 for the within county variance seems

TABLE 4-I.- ANALYSIS OF VARIANCE - WHEAT PROPORTION PER
1-SQUARE-MILE AREA (SRS SEGMENT) FOR FIVE INTENSIVE
TEST SITES IN KANSAS

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Total	86	3.242011	0.037698
LACIE segment	4	.930369	.232592
SRS segment	82	2.311642	.028191

large, and could be due to employing only five segments, three only partially available, located in five different counties in the state. Had there been more segments available, it could be that a smaller and better estimate of the variance would have resulted.

4.3 VARIANCE COMPONENTS

Utilizing the CAMS computed wheat proportions for the 28 segments processed during the November/December 1974 operation, an estimate of LACIE system variance was obtained of 0.0093. (See ref. 4 for details on its computation.) Since this is an estimate of the sum of two variance components, one due to classification errors and the other due to sampling error, and the sampling error variance component is estimated to be 0.00204 in the previous section, an estimate of the classification error variance component is $0.0093 - 0.00204 = 0.00726$. This is almost 3.6 (i.e., $0.00726/0.00204$ times as big as the sampling error component of the total error).

Thus, the classification component of the variability in estimating the wheat proportion in a county dominates heavily the sampling error component, even when only one segment per county is selected. It is conjectured that this trend in their relative magnitudes will continue to increase when more and more segments are aggregated for a crop acreage estimate.

If this is true, it implies that the classification component of the variance in acreage estimation will become more and more critical compared to the sampling error component as the level of aggregation goes up from the CRD level. Therefore, if the LACIE 90/90 criteria is not met at the desired aggregation level under the current sample segment allocation, additional sampling of segments may help in achieving it (the 90/90 criteria). This assumes that the ADP classification component of the variance does not show any significant increase relative to the sampling error component, which will be reduced by sampling of more segments.

SECTION 5.0

REFERENCES

1. Accuracy Assessment Plan for the November-December Operations. Submitted on March 19, 1975.
2. CAMS Accuracy Assessment Report for the LACIE November-December Operation. Technical Report, LEC-6060, April 1975. (JSC Rpt. No. 09710)
3. Chhikara, R. S.: Effect of Analyst Misinterpretation of Training Data on Classification Performance. Technical Report, LEC-4353, March 1975.
4. Chhikara, R. S.: LACIE - Accuracy Assessment - November/December Operations, Technical Memorandum, LEC 6037, April 1975.
5. Cochran, W. G.: Sampling Techniques. Wiley (New York), 1963.
6. Raj, Desh: Sampling Theory. McGraw-Hill (New York), 1968.
7. Warren, Fred: Variability of the Proportion of Wheat Within and Between Kansas Counties as Related to LACIE. Communication, USDA LACIE Project Office, Washington, D.C.

APPENDIX A

Model

The field data structure generated by CAMS operation is appropriately described by the model:

$$Y_{ijkl} = A + S_i + B_j + (SB)_{ij} + P_{k(i)} + (BP)_{jk(i)} + \beta X_{ijkl} + \epsilon_{ijkl}$$

$i = 1, 2, 3$ (segment number)

$j = 1, 2, 3, 4$ (biophase number)

$k = 1, 2$ (AI number)

$l = 1, 2, \dots, N_{ij}$ (field number)

where

Y_{ijkl} = LACIE-computed acreage corresponding to the l th test field of the i th segment for k th photointerpretation of j th biophase.

X_{ijkl} = Actual size of l th test field of i th segment for k th photointerpretation of j th biophase.

A = Overall bias at $X = 0$.

S_i = i th segment effect on acreage performance.

B_j = j th biophase effect on acreage performance.

$(SB)_{ij}$ = Effect of interaction between i th segment
and j th biophase on acreage performance.

$P_{k(i)}$ = k th photointerpretation effect on acreage
performance in i th segment.

$(BP)_{jk(i)}$ = Effect of interaction between j th biophase
and k th photointerpretation on acreage per-
formance in i th segment.

ϵ_{ijkl} = Error term associated with Y_{ijkl} .

The basis for this model is a cross classification between segments and biophases but a nested classification for AI's in segments. The analysis of variance obtained under this model for wheat and nonwheat field data is shown in tables A-I and A-II. Numerical values of coefficients of different terms in the model are given in table A-III.

Since the field size varies and may cause a larger (or smaller) residual, corresponding to a large (or small) field, the model needs to be checked for its adequacy. Except for a slight trend, residuals were found well scattered and random. To investigate it further, we analyzed the field data somewhat differently: we computed p , the proportion of correctly classified pixels per test field; applied a variance stabilizing transformation,

$$Y = \text{Arc sin } \sqrt{p}$$

and performed the analysis of variance for Y's. Once again, our conclusions regarding the significance of different factors in each case, wheat or nonwheat, were the same as mentioned previously.

TABLE A-I.- ANALYSIS OF VARIANCE - WHEAT FIELD DATA

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-Ratio
Total	239	680829.163	2848.66	
Field size (covariate)	1	369541.583		
Residual (total less field size)	238	311287.580	1307.93	
Segment	2	12796.741	6398.37	0.63
PI ^a in segment	3	30255.183	10085.06	^b 13.80
Biophase	3	39925.136	13308.38	^b 5.13
Segment versus biophase	6	47855.933	7980.99	^b 3.08
PI versus biophase in segment	9	23335.268	2592.81	^b 3.55
Error	215	157119.319	730.79	

^aPhotointerpretation.

^bIt implies that the factor is significant; i.e., the system performance is significantly different for different individual elements of the factor.

TABLE A-II.- ANALYSIS OF VARIANCE - NONWHEAT FIELD DATA

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-Ratio
Total	234	682015.00	2914.60	
Field size (covariate)	1	546613.10		
Residual (total less field size)	233	135401.90	581.12	
Segment	2	23124.605	11562.30	^b 26.18
PI ^a in segment	3	1324.967	441.65	.95
Biophase	3	5673.291	1891.09	^b 6.93
Segment versus biophase	6	5395.846	899.30	^b 3.30
PI versus biophase in segment	9	2454.592	272.73	.59
Error	210	97428.601	463.95	

^a Photointerpretation.

^b It implies that the factor is significant; i.e., the system performance is significantly different for different individual elements of the factor.

TABLE A-III.- MAIN COEFFICIENTS FOR THE MODEL

Term in model	Coefficient value	
	Wheat	Nonwheat
β	0.646	0.924
A	0.54	-6.5
S_1	10.5	3.0
S_2	-15.4	14.1
S_3	4.9	-17.1
B_1	-21.9	6.7
B_2	10.7	-0.9
B_3	3.9	-7.1
B_4	7.3	1.3

TABLE A-IV.- ITS WHEAT PERCENTAGE DATA

ITS segment	Biophase	Wheat percentage	
		Actual	LACIE ^a
1034 (Finney)	I	23	17.49
			13.39
	II		43.90
			27.14
	III		24.48
1042 (Norton)		40	21.00
	IV		37.12
			22.98
	I		37.33
			40.52
1106 (Ellis)	II	45	41.46
			26.20
	IV		31.37
			22.00
	I		39.37
1111 (Rice)		34	59.06
	II		68.78
			50.96
	III		60.77
			73.09
1114 (Saline)	IV	59	39.97
			72.15
	III		17.55
	IV		28.89
			49.37
	II		65.07
	II		44.37
			40.87
	IV		53.53
			39.42

^aTwo LACIE estimates corresponding to each biophase due to two photointerpretations.

TABLE A-V.- ANALYSIS OF VARIANCE - ITS WHEAT PERCENTAGES

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-Ratio
Total	29	8309.54	276.74	
Actual percentage (covariate)	1	2211.91		
Residual (total less covariate)	28	6097.73	217.78	
Biophase	3	125.81	41.94	0.18
Error	25	5971.92	238.88	

TABLE A-VI.- COEFFICIENTS IN THE MODEL^a

Term in model	Coefficient value
β	0.74
A	11.79
B_1	-2.48
B_2	1.64
B_3	-1.34
B_4	2.18

^aModel: $Y_{ijk} = A + B_j + \beta X_i + \epsilon_{ijk}$
 $i = 1, 2, 3, 4, 5$ (segment number)
 $j = 1, 2, 3, 4$ (biophase number)
 $k = 1, 2$ (photointerpretation number)

where

Y_{ijk} = LACIE-computed wheat percentage for ITS area in
 i th segment with inputs from j th biophase and
its k th photointerpretation.

X_i = Actual wheat percentage for ITS area in i th segment.

A = Overall bias at $X = 0$.

B_j = Effect of j th biophase on wheat estimation.

ϵ_{ijk} = Error term associated with Y_{ijk} .

APPENDIX B
INTRA-CLUSTER CORRELATION COEFFICIENT

Formula

Let y_{ij} be the measurement for the j th individual of the i th cluster, $j = 1, 2, \dots, M$ and $i = 1, 2, \dots, N$, and $\bar{y}_i = \sum_j y_{ij}/M$, the i th cluster average, and $\bar{y} = \sum_{ij} y_{ij}/NM$, the overall average. Then, by definition, the intra-cluster correlation is

$$r = \frac{\sum_{ijk} (y_{ij} - \bar{y}_i) (y_{ik} - \bar{y}_i) / NM(M - 1)}{\sum_{ij} (y_{ij} - \bar{y})^2 / MN}$$

This simplifies to (e.g., see ref. 6, p. 110)

$$r = \frac{M \sigma_b^2 - \sigma^2}{(M - 1) \sigma^2}$$

where

σ^2 = Total (between cluster plus within cluster) variance.

σ_b^2 = Between cluster variance component.

M = Number of individuals per cluster.

Calculations

From table 3-I, estimates corresponding to σ^2 and σ_b^2 are given by

$$s^2 = 0.037698$$

and

$$s_b^2 = \frac{0.232592 - 0.028191}{30} = 0.006813$$

respectively. So, in the present case,

$$r = \frac{30(0.006813) - 0.037698}{29(0.037698)} = 0.15$$

It may be mentioned that $s^2 = 0.037698$ is also an estimate of the within county variance for the SRS segment and is certainly higher than expected. This is partly due to the fact that the ITS segments were located in different counties, and so it also includes the between county variance, as well. But then, so does s_b^2 . When r is computed as above, it is hoped that the effect of the between county variance is minimized.